



Data Science in Action #3

Getting More Insight into Your Forecast Errors using Multivariate Statistics



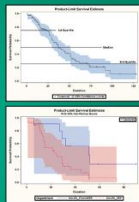
Gerhard Svolba
Data Scientist, SAS Austria

Data Science Applications and Case Studies

Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention

*Can assumptions about the average
length of time intervals be made, even if
most of the endpoints have not yet been
observed?*



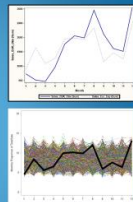
Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining



Data Science in Action: #5

Checking the Alignment with Predefined Pattern

*Which customers show a behavior that
is far from what you expected?*



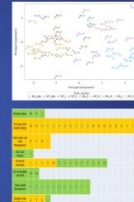
Chi2 independency test
Benford's law
Time Series Similarity



Data Science in Action: #7

Topic Search Documents and Clustering

*Can I automatically find clusters of
documents with similar content?*



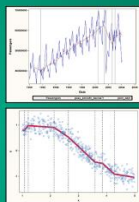
Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights



Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

*Can events and changes in the
course over time be
automatically detected?*



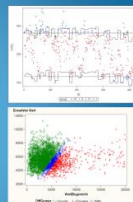
Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models



Data Science in Action: #6

Proving a reference value that considers all available co-information

*Can analytics help me to reduce the
"Yes, but ..." sentences in my business
discussions?*



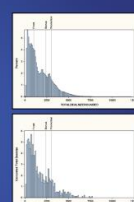
Linear Regression
Decision Trees
Time Series Analysis



Data Science in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

*When the sales manager looks at the
project pipeline, does the sum of weighted
averages give him or her a full picture?*



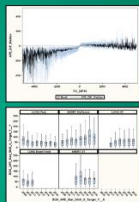
Monte Carlo Simulations
Mathematical Programming



Data Science in Action: #3

Explaining Forecast Errors and Deviations

*Do the demand planners really improve
forecast accuracy with their manual
overwrites?*



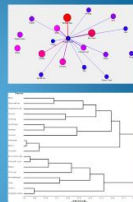
Linear Regression
Quantile Regression
Descriptive Statistics



Data Science in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

*Can your data tell you stories about
your analysis subjects, even if you don't
ask explicitly?*



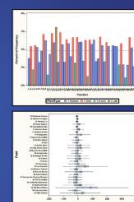
Unsupervised machine learning methods:
association analysis
variable clustering



Data Science in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

*How can you simulate complex
environments to get insight in the most
frequent processes?*



Monte Carlo Simulations

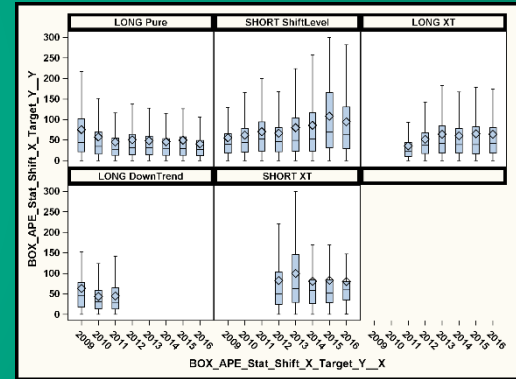
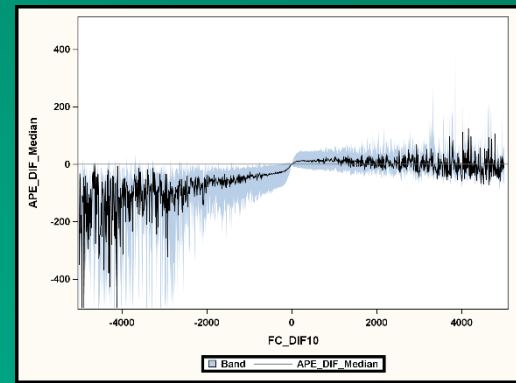


Data Science in Action: #3

Explaining Forecast Errors and Deviations

*Do the demand planners really improve
forecast accuracy with their manual
overwrites?*

Linear Regression
Quantile Regression
Descriptive Statistics



Paper SAS1673-2018 Getting More Insight into Your Forecast Errors with the GLMSELECT and QUANTSELECT Procedures

Gerhard Svolba, SAS Institute Inc. Austria
Denver, April 10th, 2018

Twitter: @gsvolba



USERS PROGRAM

SAS GLOBAL FORUM 2018



This Presentation Provides You

#SASGF



Analytic
Business Questions



Advanced Analytic
Methods



Business Results
and Actions



6 Relevant
Graph Examples



10 SAS Code
Snippets



Data Preparation
Considerations



Regression Procedures
in SAS Viya

SAS GLOBAL FORUM 2018

SAS Global Forum Paper SAS1673-2018 Getting More Insight into Your Forecast Errors with the GLMSELECT and QUANTSELECT Procedures

Gerhard Svolba, SAS Institute Inc. Austria

Denver, April 10th, 2018



Paper SAS1673-2018
Getting More Insight into Your Forecast Errors with the GLMSELECT and
QUANTSELECT Procedures
Gerhard Svolba, SAS Institute Inc. Austria

ABSTRACT
In a forecast you monitor the quality of your forecast models over time? Can data science methods identify the areas for large forecast errors and provide more insights than descriptive statistics? Do demand planners really improve forecast accuracy with their neural networks? Using a real life case study, this paper answers these questions. It shows you how to use the power of SAS to identify the product, group, forecast horizon, seasonality, or the forecast model type on forecast accuracy and current facts on alternative models. You learn how to apply methods provided for length into the evaluation and relationships of your forecast data. You gain insight into how manual correction of the statistical forecast change forecast accuracy in both directions and how you can apply statistical and graphical methods to achieve more insight. The paper shows you how to use the GLMSELECT and QUANTSELECT procedures to provide additional relevant insight. You learn how to use the GLMSELECT and QUANTSELECT procedures to identify the most important relationships in the forecast error. You see how you can enhance and interpret the output of these procedures to quantify the effect of the relevant factors. You learn how to convert the results from the GLMSELECT and QUANTSELECT procedures into actionable forecast insights. The paper shows a number of how to use the REGSELECT and QTRSELECT procedures to apply these methods in SAS Viya.

INTRODUCTION
APPLY ANALYTICAL METHODS ACROSS DIFFERENT BUSINESS DOMAINS
Analytical methods can leverage the analysis solutions for various business questions. Using one level of analysis, you can identify and analyze forecast errors in the relationship between different categories. Analytical methods also help you spot relationships and enable you to receive an objective and data driven answer to your business questions.
The book Applying Data Science: Business Case Studies Using SAS® (SAS 2017) is dedicated to the application of analytical methods to different types of business questions. It shows how analytical methods that have been successfully used in certain business domains can and should be applied also to other business areas. For example, you can apply several analysis techniques to analyze the relationship between all categories, or you can use various statistical and machine learning regression options to automatically select variables in your time series data.

CASE STUDY: ANALYZING THE FORECAST ERROR
This paper looks at a case study from the demand forecasting area. The focus is to investigate the forecast error, which is measured as the deviation between the forecasted demand and the actual demand. It shows how analytical methods can be used to identify factors that influence the forecast error. The case study does not deal with the creation of the statistical forecast itself but with the evaluation of the forecast quality. Several methods are used to evaluate the forecast quality and the paper shows how they can be applied with analytical methods, like descriptive statistics and various linear models.

STATISTICAL REGRESSION ANALYSIS
The statistical data that are shown here include location, histogram, and descriptive measures like mean, median, and quartiles, as well as linear regression and quantile regression models.

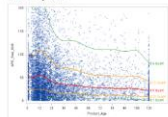
Step	Efficientest	MAIC
1	Model	43544.0038
2	Learn_Metric	43116.2647
3	Learn_Time	43387.2543
4	Learn_Group	43494.7532
5	Product_Group	43077.2763
6	Product_Group	43123.0880
7	Product_Group	43125.2832

Table 5. Variables Selected for the 0.75 Quantile

Step	Efficientest	MAIC
1	Model	43577.4745
2	Learn_Metric	44027.8535
3	Learn_Metric	44564.7538
4	Learn_Time	43078.0409
5	Learn_Group	43078.0409
6	Product_Group	43078.0409
7	Product_Group	43078.0409
8	Product_Group	43078.0409

Table 6. Variables Selected for the 0.75 Quantile

Displaying the Results Visually
Output 4 displays the results of the multivariate quantile regression in the same way as shown in Output 3 for the 0.5 regression.

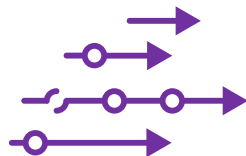


Output 4. Plot of the Predicted APE_STAT Values from the Multivariate Quantile Regression Model
The four lines represent a value for the PRODUCT, AGE, and APE_STAT of the actual data. The two confidence intervals for the predicted values of APE_STAT from the multivariate quantile regression are shown in the background. The regression of product age, however, is not only evaluated on its own, it is corrected for the effect of the other available variables as multivariate regression model is used.

This Presentation Provides You



Analytic
Business Questions



Advanced Analytic
Methods



Business Results
and Actions



Data Preparation
Considerations



6 Relevant
Graph Examples



10 SAS Code
Snippets



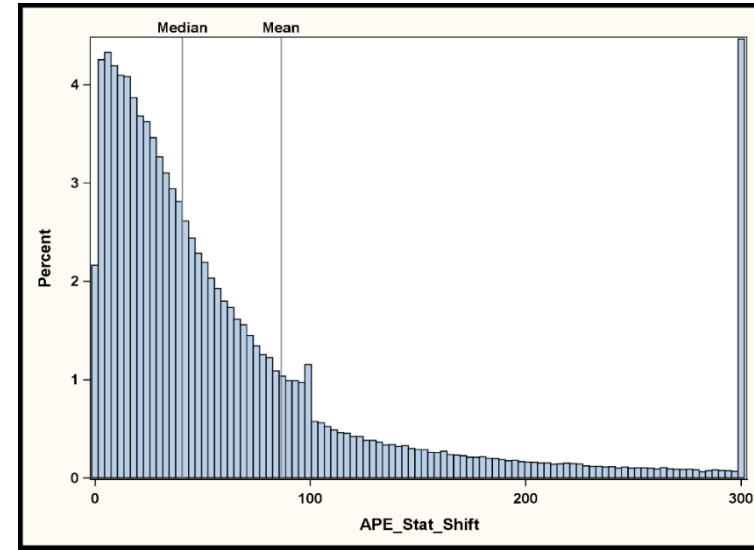
Regression Procedures
in SAS Viya

Business Background of the Case Study

- International retail and manufacturing company
- Demand forecasts on a monthly basis
- Forecasts generated
 - Long history products (>15 months) → SAS® Visual Forecasting
 - Short history (fashion) products → SAS® VDMML
- Want to understand deviation in forecast quality

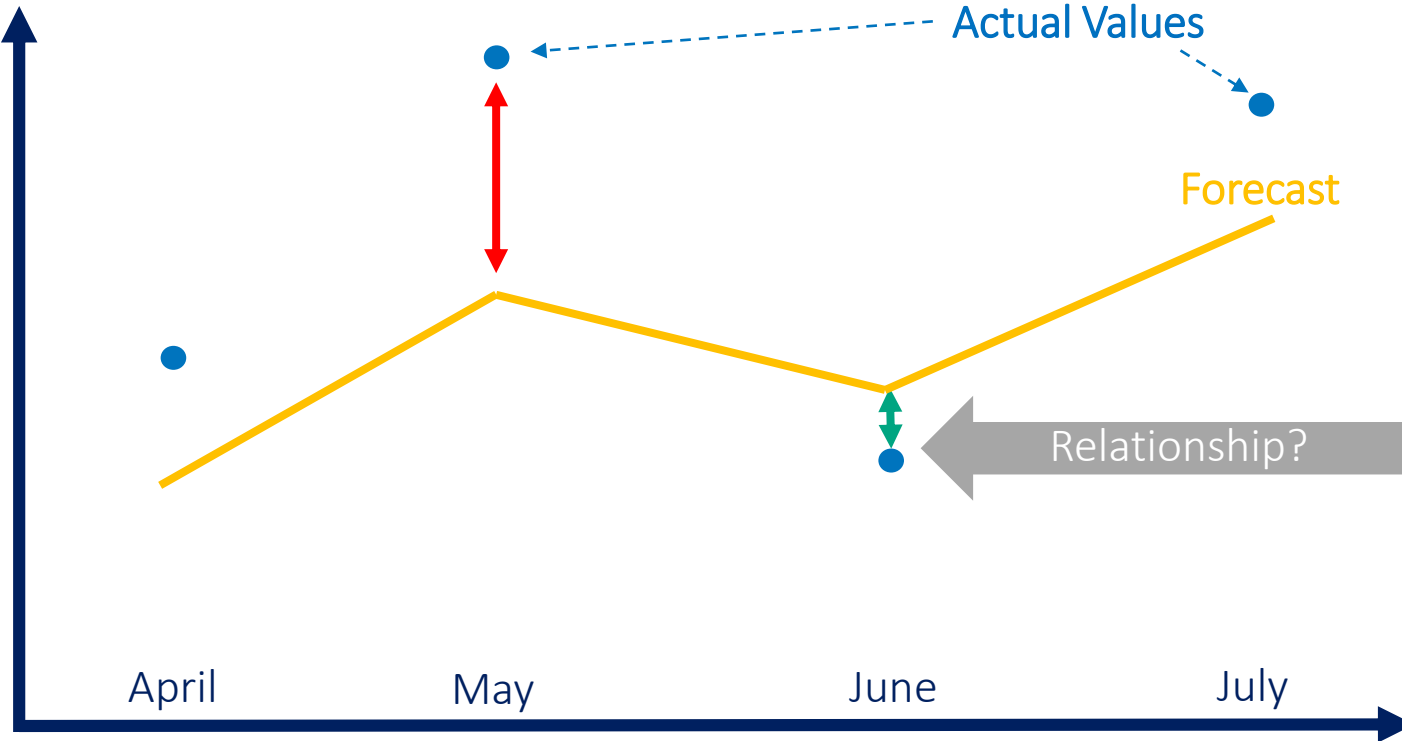
Business Questions for the Case Study

- What is the distribution of the forecast error?
- Which factors influence the forecast error?
- Where should you invest time to improve forecast quality?
- Which combinations might always have large forecast errors?
- Do manual overrides improve forecast quality?



Basic Idea: Explain the „Size“ of the Forecast Errors

Forecast for Item 1673: „GPS Tracker Waterproof“



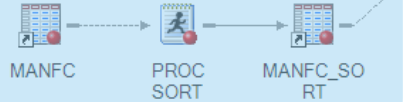
Product Group
Price
Launch Calendar Month
Product Age (=Data History)
Forecast Model
Lead Time
Target Year
Target Calendar Month

Available Data and Data Preparation

Statistical Forecast



Manual Override



Forecast Model
Lead Time
Target Year
Target Calendar Month

Product Group
Price
Launch Calendar Month
Product Age (=Data History)

$$\text{APE_STAT} = \text{abs}(\text{statfc} - \text{actual}) / \text{actual} * 100$$

→ Absolute Percentage Error

Using the MAPE

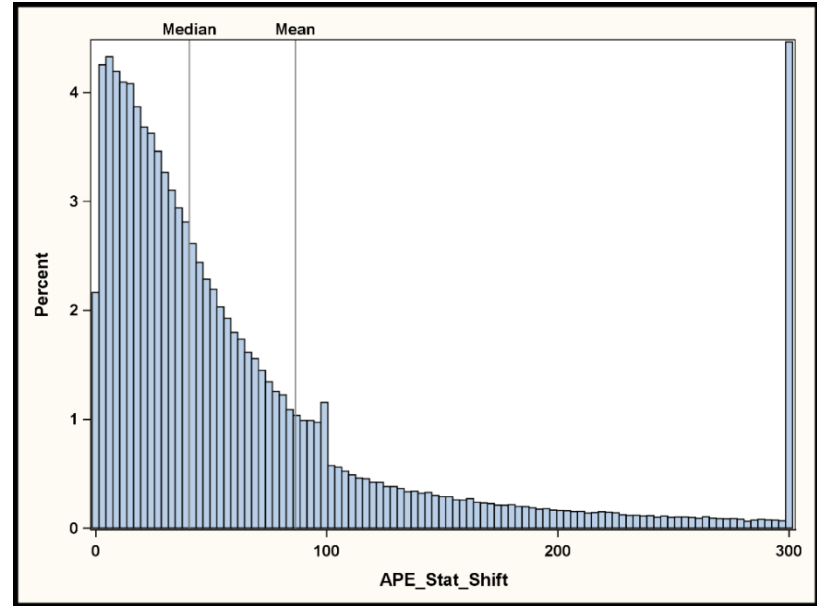
MAPE – Mean Absolute Percentage Error

- Why you might not use it:
 - MAPE is asymmetric; perfect fit results in a MAPE of 0.
 - If observed demand = 0 → MAPE formula: division by zero.
 - Forecast of 0 → MAPE=100. Forecasting might limit its forecast error by forecasting 0 for all time points.
- However:
 - INTERPRETABILITY!
 - Widely Used in Business Forecasting

Overall Distribution of the Forecast Error

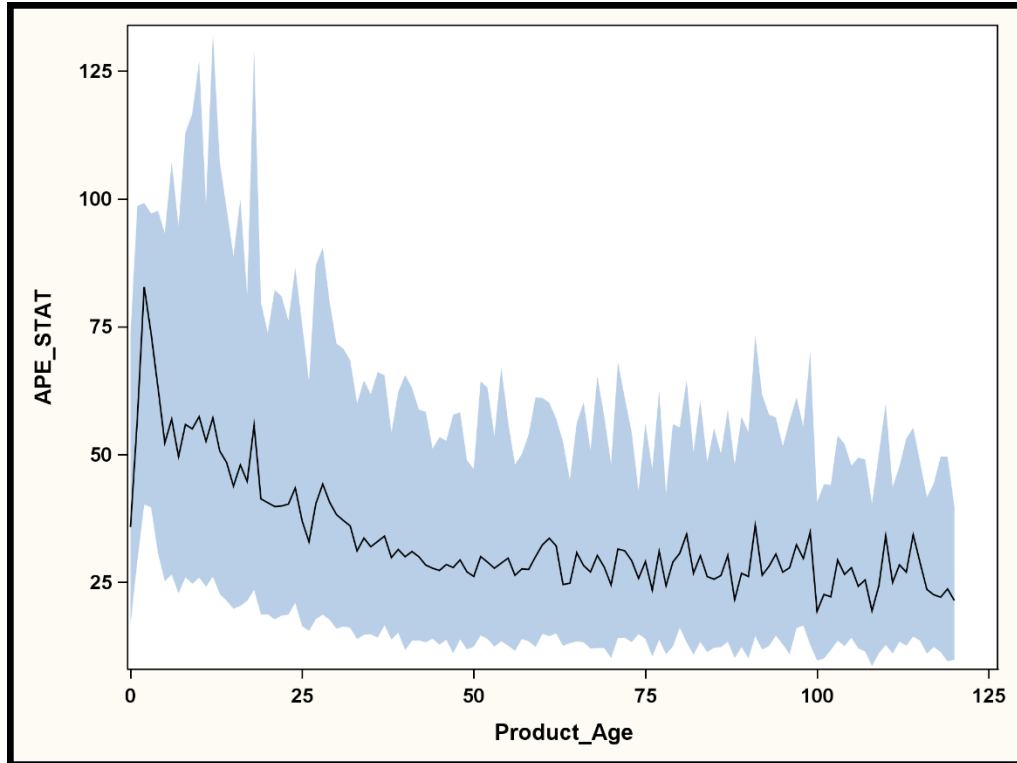
Mean of all APEs = 85.5

Quantile	Value
100% Max	238,954.6
95%	276.6
90%	169.5
75% Q3	81.7
50% Median	40.6
25% Q1	18.0
10%	7.0
0% Min	0



Using a Band (1st+3rd Quartile) and a Line (Median) Chart

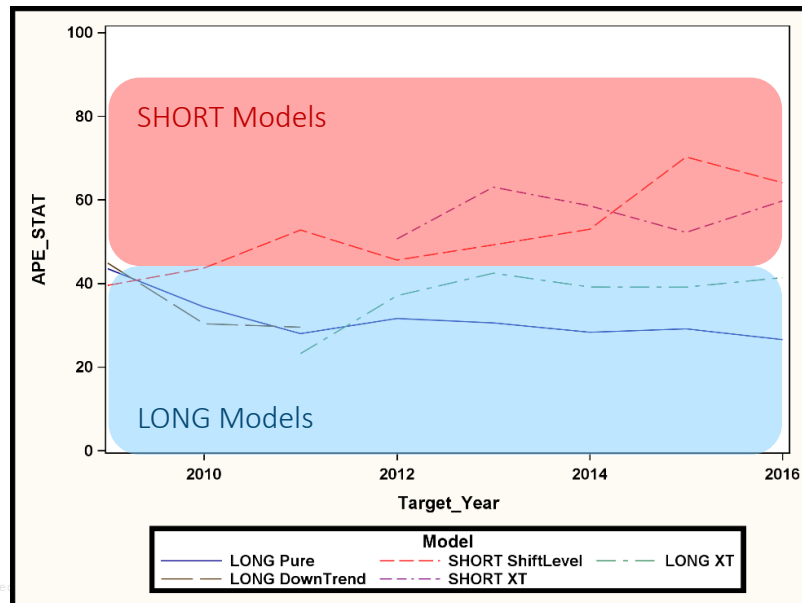
→ Longer Data History reduces Forecast Error



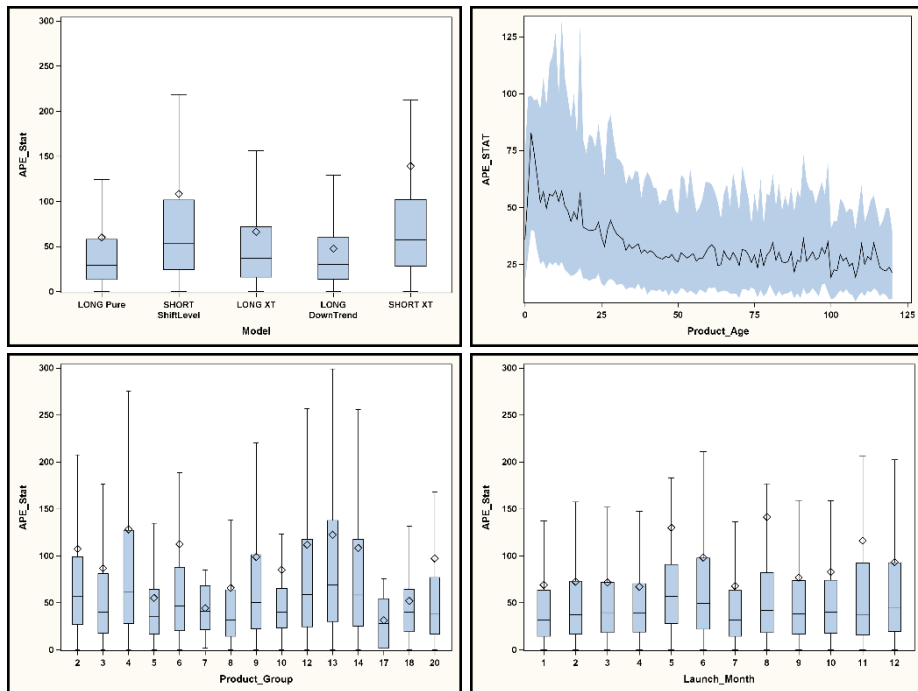
Also Descriptive Methods help!

Analyze the Forecast Error Over Time

- Forecast Errors for short-term products are higher (and increasing over the years)
- Some Forecast Models are discontinued and replaced by other Models
- Some models might exhibit a larger forecast error because they are used to forecast „special“ products



Results from Univariate Linear Regression Models



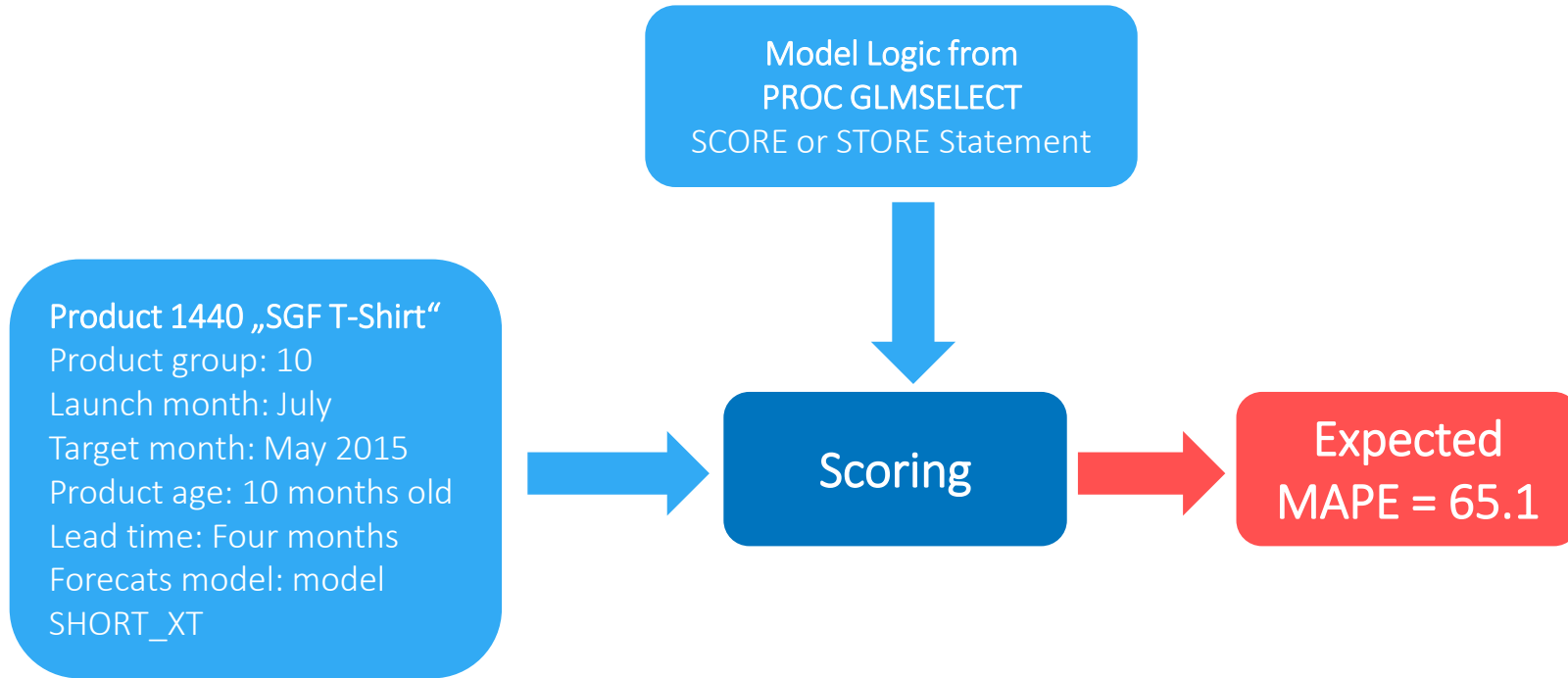
Rank	Input Variable	R-squared
1	MODEL	0.0554
2	PRODUCT_AGE	0.0433
3	PRODUCT_GROUP	0.0224
4	LAUNCH_MONTH	0.0172
5	TARGET_YEAR	0.0102
6	TARGET_CALMONTH	0.0084
7	LEAD_TIME	0.0046
8	PRICE_INDEX	0.0016

```
PROC GLMSELECT DATA=fc_mart;
  MODEL ape_stat_shift = product_Age;
RUN;
```

Comparing the Selection Order of Variables in the Univariate and the Multivariate Linear Regression Model

Rank	Input Variable	Adjusted R-square	Beta (Good/Bad)	Rank (Change)
1	MODEL	5.46%	Long Short	1 (=)

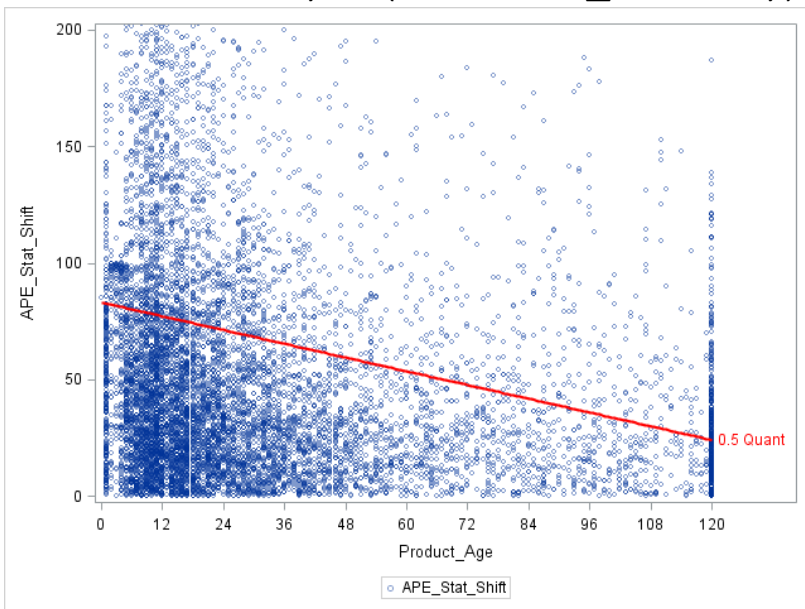
Use the Regression Model to Calculate the expected MAPE for new Data



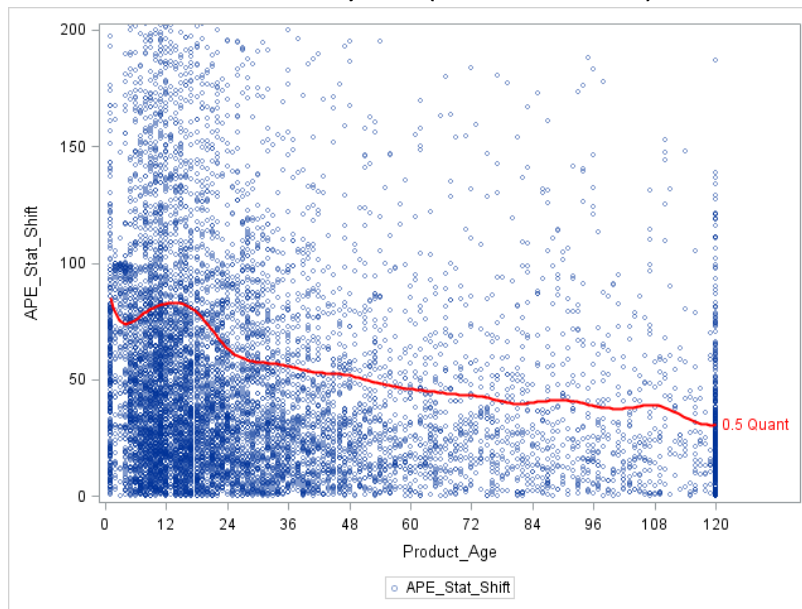
Studying Linear Regression Results Visually

Influence of Variable PRODUCT_AGE

Univariate Analysis (PRODUCT_AGE only)

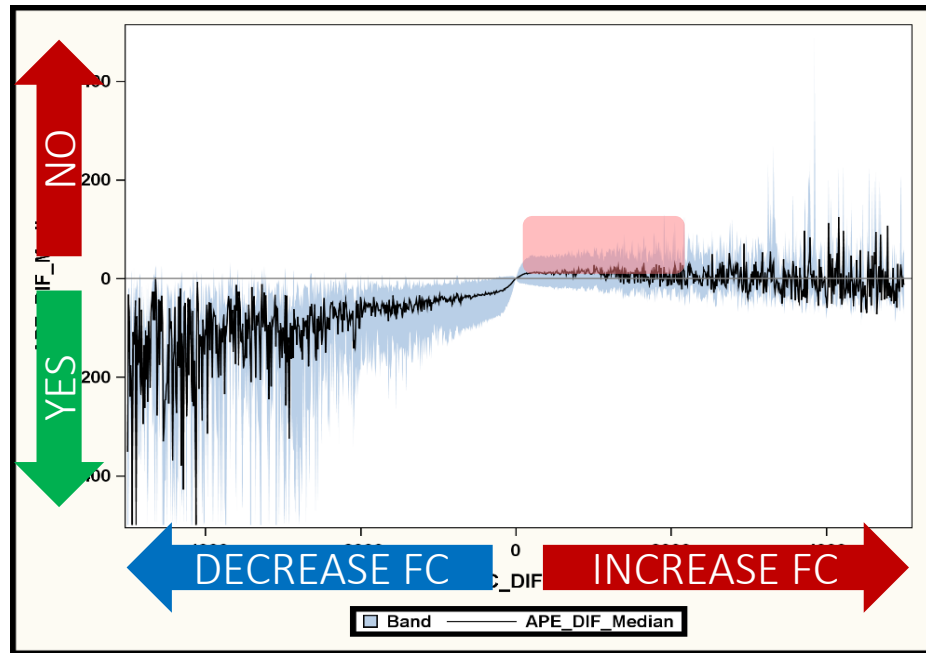


Multivariate Analysis (8 variables)



Do Demand Planners improve Forecast Quality with their Manual Overrides?

- Line and Band Chart:
 - The median is shown by a solid black line.
 - The first and third quartile are displayed by a band.
- Larger changes → Larger effect
- Corresponds with the work of Paul Goodwin (2009)
- Demand planners obviously put more thought into large changes 😊
- Eliminate the small changes in your process! (Usually do not add any benefit.)



Possible next Steps

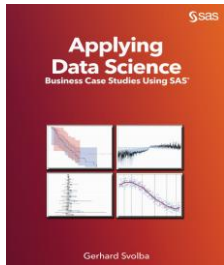
- Build a decision tree to discover segments with high/low forecast error
- Build a machine learning model that calibrates/suggests the optimal override
 - FVA (Forecast Value Add) Analysis
 - Also consider additional explanatory variables (product and forecast features)

Take-Aways from this Presentation

- **Application of analytical methods** provides relevant insights and help you make better business decisions.
- **Descriptive and visual methods** also provide a lot of insight to understand business relationships
- **Multivariate regression analysis** provides a more comprehensive picture than the isolated univariate analysis of influential factors.
- **Quantile regression** enables you get a clearer picture about the extremes of your distribution.
- The **SAS platform with SAS9 and SAS Viya** procedures provides a comprehensive set of analytical methods

Analytics and Data Science is there to help you!

- Get a clearer, more objective picture of your data and your analysis subjects
- Get explicit results instead of searching the needle in the haystack
- Make your data talk to you!
- Receive findings automatically instead of manually
- Do it again! – treat models as an asset and repeat your analysis



Get access to more content:



SAS DACH @Youtube: <https://www.youtube.com/user/SASsoftwareGermany>

Blogs on LinkedIn: <https://www.linkedin.com/in/gerhardsvolba/>

Twitter: <https://twitter.com/gsvolba>

Content on Github: <https://github.com/gerhard1050>

Books @SAS-Press: <https://support.sas.com/svolba>

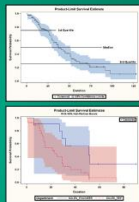


Data Science Applications and Case Studies

Data Science in Action: #1

Performing Headcount Survival Analysis for Employee Retention

*Can assumptions about the average
length of time intervals be made, even if
most of the endpoints have not yet been
observed?*



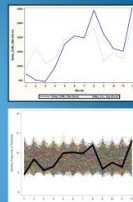
Survival analysis methods: Kaplan-Meier estimates
Cox Proportional Hazards regression
Survival Data Mining



Data Science in Action: #5

Checking the Alignment with Predefined Pattern

*Which customers show a behavior that
is far from what you expected?*



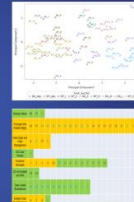
Chi2 independency test
Benford's law
Time Series Similarity



Data Science in Action: #7

Topic Search Documents and Clustering

*Can I automatically find clusters of
documents with similar content?*



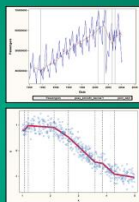
Text Mining
Text Parsing (Synonyme, Stemming, Stop-Listen)
Term by Document Weights



Data Science in Action: #2

Detecting Structural Changes and Outliers in Longitudinal Data

*Can events and changes in the
course over time be
automatically detected?*



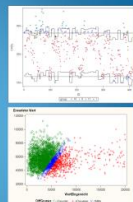
Smoothing Of Longitudinal Data
Multivariate Adaptive Regression Splines
Automatic Breakpoint Detection
Automatic Detection of Outliers with ARIMA Models



Data Science in Action: #6

Proving a reference value that considers all available co-information

*Can analytics help me to reduce the
"Yes, but ..." sentences in my business
discussions?*



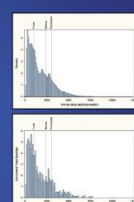
Linear Regression
Decision Trees
Time Series Analysis



Data Science in Action: #8

Using Monte Carlo Simulations to Understand the Outcome Distribution

*When the sales manager looks at the
project pipeline, does the sum of weighted
averages give him or her a full picture?*



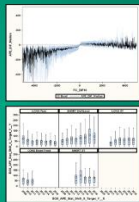
Monte Carlo Simulations
Mathematical Programming



Data Science in Action: #3

Explaining Forecast Errors and Deviations

*Do the demand planners really improve
forecast accuracy with their manual
overwrites?*



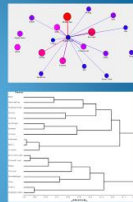
Linear Regression
Quantile Regression
Descriptive Statistics



Data Science in Action: #4

Listening to Your Data – Discover Relationships with Unsupervised Analysis Methods

*Can your data tell you stories about
your analysis subjects, even if you don't
ask explicitly?*



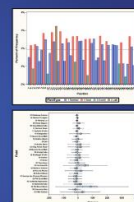
Unsupervised machine learning methods:
association analysis
variable clustering



Data Science in Action: #9

Studying Complex Systems – Simulating the Monopoly Board Game

*How can you simulate complex
environments to get insight in the most
frequent processes?*



Monte Carlo Simulations



